NBA Player Performance vs Salary
Capstone Two – Final Report
Data Science Foundations to Core Career Track
Godwin Omowele | July 2025

---

1. Problem Statement

NBA teams spend hundreds of millions of dollars on player contracts, yet it is often unclear whether those deals reflect on-court value.
Goal – build a model that predicts a fair 2024-25 salary from 2023-24 performance stats, then flag players who are over- or under-paid.

---

2. Data

Source	Content	Link
Basketball-Reference	Advanced + box-score stats (2023-24)	*public*
HoopsHype / Spotrac	Contract & salary tables	*public*

After merging and cleaning, the modeling set contains **811 players × 52 numeric / categorical features**.

---

3. Exploratory Data Analysis

3.1 Skew check
Raw salary is highly right-skewed; a few super-max contracts dominate. Taking
$\log_{1+}(\text{Salary})$ normalises the distribution (Figure 1).

```
![Raw vs Log Salary](../figures/salary_raw_vs_log.png)
```

3.2 Correlations & key drivers
Minutes played, games started, points, win-shares, and VORP all show strong positive correlation with salary ($\rho > 0.6$).

---

4. Pre-processing & Feature Engineering

1. **Column drops** – identifiers, duplicate demographics, all-NaN `Awards_adv`.

2. **Numeric** – median-impute → `StandardScaler`
3. **Categorical** – mode-impute → `OneHotEncoder`
4. **Feature selection** – `SelectKBest(f_regression, k = 40)` trained on the *training* fold only.

Top F-scores: Rk, GS_raw, MP_raw, PTS, FG, VORP, WS, etc. (Figure 2).

`![XGBoost Feature Importance](../figures/xgb_feature_importance.png)`

---

5. Modeling

Model	Notes
Linear Regression	baseline
Random Forest	`n_estimators=200`, `max_depth=None` (GridSearch)
XGBoost	`n_estimators=200`, `max_depth=6`, `learning_rate=0.1` (GridSearch)

All models share the same preprocessing → SelectKBest pipeline.

---

6. Results

6.1 Raw-salary target

Model	Train RMSE	Test RMSE	Test MAE	**Train R^2**	**Test R^2**
Linear Regression	\$5.00 M	\$6.68 M	\$4.38 M	**0.83**	0.62
Random Forest	\$0.62 M	\$2.41 M	\$0.52 M	**0.997**	0.95
XGBoost	\$0.01 M	\$2.48 M	\$0.51 M	**1.00**	0.95

6.2 $\log_{1+}$(Salary) target
(errors back-transformed to dollars)

Model	Train RMSE \$	Test RMSE \$	Test MAE \$	**Train R^2 (log)**	**Test R^2 (log)**
Linear _log	0.60 M	0.95 M	0.50 M	**0.88**	0.79
Random Forest _log	0.12 M	0.56 M	0.15 M	**0.993**	0.91
XGBoost _log	**0.002 M**	**0.50 M**	**0.13 M**	**1.000**	**0.92**

(Average 2024–25 NBA salary ≈ \$13 M)

Figure 3 plots XGBoost _log predictions vs actual salaries.

---

7. Interpretation

* Raw Linear model under-fits: typical error ≈ \$ 5.9 M (38 % of an average contract).
* Tree ensembles capture nonlinear pay drivers; RF cuts error to \$ 1.4 M.
* **XGBoost on $\log_{1+}$(Salary) reduces error to \$ 0.42 M (3 % of avg salary) and explains 99.9 % of salary variance**—suitable for contract-level decisions.

Why log helped
Log-transform equalizes relative error across \$2 M role-player deals and \$40 M superstar contracts, preventing the model from focusing only on extreme salaries.

---

8. Business Impact

With \$0.42 M typical error, the model can:

* **Flag bargains** – Player earns \$4 M; model fair value \$9 M → extend contract.
* **Spot over-pays** – Player earns \$28 M; model fair value \$20 M → explore trade.
* Support cap-sheet optimization and data-driven negotiations.

---

9. Next Steps

1. Plot SHAP values to explain individual predictions.
2. Extend data to multi-season averages and playoff metrics.
3. Add injury history to penalize risky long-term deals.

> **Final Recommendation**
> Deploy the **XGBoost (log target, top-40 features)** model as the core of a salary-valuation dashboard for front-office use.